DOCUMENT RESUME

ED 316 580 TM 014 551

AUTHOR Rosenthal, Robert
TITLE Research: How Are We Doing?
SPONS AGENCY National Science Foundation, Washington, D.C.
PUB DATE Apr 89
NOTE 55p.; Paper presented at the Annual Meeting of the Eastern Psychological Association (Boston, MA, March 30-April 2, 1989).
PUB TYPE Information Analyses (070) -- Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Effect Size; Mathematical Models; Meta Analysis; *Psychological Studies; *Research Methodology; Sampling; Statistical Bias; Statistical Significance
IDENTIFIERS Binomial Effect Size Display; *Research Replication; Type II Errors

ABSTRACT

An overview of the state of the art in psychological research is presented, with an emphasis on the attention given to effect sizes. The acceptance of small effect sizes for biomedical research is contrasted with the rejection of similar effect sizes for psychological research. The Binomial Effect Size Display is used to depict the practical magnitude of an effect size regardless of whether the dependent variable is dichotomous or continuous. Other topics discussed include: (1) the meaning of successful replication, including successful replication of Type II errors; (2) reporting results of replications, including tests of significance; (3) meta-analytic procedures; (4) sampling bias; (5) overemphasis on single values and disregard of details; (6) problems of heterogeneity of method and quality; (7) problems of independence of responses within a single study and within sets of studies; and (8) exaggeration of significance levels. Several benefits of meta-analysis are outlined. It is concluded that many findings of psychological research are neither small nor practically unimportant. Nevertheless, it is also concluded that in the areas of replication and of the cumulation of research findings much remains to be done. Eight data tables and one graph are provided. A 46-item list of references is included. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED316580

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Robert Rosenthal
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Research: How Are We Doing?

Robert Rosenthal

Harvard University

My talk today is designed in part both to comfort the afflicted and to afflict the comfortable. The afflicted are those of us who work in the softer, wilder areas of our field--the areas where the results seem ephemeral and unreplicable, and where the r^2's seem always to be approaching zero as a limit. These softer, wilder areas include those of social, personality, clinical, developmental, educational, organizational, and health psychology. They also include parts of psychobiology and cognitive psychology. These softer, wilder areas, however, may not include too much of psychophysics.

My message to those of us toiling in these muddy vineyards will be that we are doing better that we might have thought. My message to those of us in any areas in which we feel we have pretty well nailed things down will be that we haven't, and that we could be doing a whole lot better.

How Large Must an Effect Be, To Be Important?

There is a bit of good news-bad news abroad in the land. The good news is that more sophisticated editors, referees, and researchers are becoming aware that reporting the results of a significance test is not a sufficiently enlighteningprocedure

TM014551

BEST COPY AVAILABLE

to stand alone. More and more we are beginning to see a report of the magnitude of the effect accompanying the *p* level. The bad news is that we are still not quite sure what to do with such a report of the magnitude of the effect, for example, a correlation coefficient.

There is one bit of training that all psychologists have undergone. From undergraduate days onward we have all been taught that there is only one proper, decent thing to do whenever we see a correlation coefficient--we must square it. For most of the softer, wilder areas of psychology, squaring the correlation coefficient tends to make it go away--vanish into nothingness as it were. That is one of the sources of malaise in the social and behavioral sciences. It is sad and quite unnecessary, as we shall soon see.

The Physician's Aspirin Study

At a special meeting held on December 18, 1987, it was decided to end prematurely, a randomized double blind experiment on the effects of aspirin on reducing heart attacks (Steering Committee of the Physicians' Health Study Research Group, 1988). The reason for this unusual termination of such an experiment was that it had become so clear that aspirin prevented heart attacks (and deaths from heart attacks) that it would be unethical to continue to give half the

physician research subjects a placebo. Now what do you suppose was the magnitude of the experimental effect that was so dramatic as to call for the termination of this research? Was r^2 .90, or .80, or .70, or .60, so that the corresponding r's would have been .95, .89, .84, or .77? No. Well, was r^2 50, .40, .30, or even .20, so that the corresponding r's would have been .71, .63, .55, or .45? No. Actually, what r^2 was, was .0011, with a corresponding r of .034.

Insert Table 1 about here

Table 1 shows the results of the aspirin study in terms of raw counts, percentages, and as a Binomial Effect Size Display (BESD). This display is a way of showing the practical importance of any effect indexed by a correlation coefficient. The correlation is shown to be the simple difference in outcome rates between the experimental and the control groups in this standard table which always adds up to column totals of 100 and row totals of 100 (Rosenthal & Rubin, 1982b).

This type of result seen in the physicians' aspirin study is not at all unusual in biomedical research. Some years earlier, on October 29, 1981, the National Heart, Lung, and Blood Institute discontinued its placebo-controlled study of propranolol because results were so favorable to the treatment that it would be unethical to

continue withholding the life-saving drug from the control patients. And what was the magnitude of this effect? Once again the effect size *r* was .04, and the leading digits of the r^2 were .00! As behavioral researchers we are not used to thinking of *r*'s of .04 as reflecting effect sizes of practical importance. But when we think of an *r* of .04 as reflecting a 4% decrease in heart attacks, the interpretation given *r* in a Binomial Effect Size Display, the *r* does not appear to be quite so small; especially if we can count ourselves among the 4 per 100 who manage to survive.

Insert Table 2 about here

Additional Results

Table 2 gives three further examples of Binomial Effect Size Displays. In a recent study of 4,462 Army veterans of the Vietnam War era (1965-1971), the correlation between having served in Vietnam (rather than elsewhere) and having suffered from alcohol abuse or dependence was .07 (Centers for Disease Control, 1988). The top display of Table 2 shows that the difference between the problem rates of 53.5 and 46.5 per 100 is equal to the correlation coefficient of .07.

The center display of Table 2 shows the results of a study of the effects of AZT on the survival of 282 patients suffering from AIDS or AIDs-related complex (ARC)

(Barnes, 1986). This result of a correlation of .23 between survival and receiving AZT (an r^2 of .054) was so dramatic as to lead to the premature termination of the clinical trial on the ethical grounds that it would be improper to continue to give placebo to the control group patients.

As a footnote to this display let me add the result of a small informal poll I took a few weeks ago of some physicians spending the year at the Center for Advanced Study in the Behavioral Sciences. I asked them to tell me of some medical breakthrough that was of very great practical importance. Their consensus was that the breakthrough was the effect of cyclosporine in increasing the probability that the body would not reject an organ transplant and that the recipient patient would not die. A multi-center randomized experiment was published in 1983 (Canadian Multicentre Transplant Study Group, 1983). The results of this breakthrough experiment were less dramatic than the results of the AZT study. For the dependent variable of organ rejection the effect size r was .19 (r^2 = .036); for the dependent variable of patient survival the effect size r was .15 (r^2 = .022).

The bottom display of Table 2 shows the results of a famous meta-analysis of psychotherapy outcome studies reported by Smith and Glass (1977). An eminent critic (Rimland, 1979) believed that the results of their analysis sounded the "death

knell" for psychotherapy because of the modest size of the effect. This modest effect size was an *r* of .32 accounting for "only 10% of the variance."

Examination of the bottom display of Table 2 shows that it is not very realistic to label as "modest indeed" an effect size equivalent to increasing a success rate from 34% to 66% (for example, reducing a death rate or a failure rate from 66% to 34%). Indeed, as we have seen, the dramatic effects of AZT were substantially smaller (r = .23), and the "breakthrough" effects of cyclosporine were smaller still (r = .19).

Telling How Well We're Doing

The Binomial Effect Size Display is a useful way to display the practical magnitude of an effect size regardless of whether the dependent variable is dichotomous or continuous (Rosenthal & Rubin, 1982b). An especially useful feature of the display is how easily we can go from the display to an *r* (just take the difference between the success rates of the experimental versus the control group) and how easily we can go from an effect size *r* to the display (just compute the treatment success rate as .50 plus one-half of *r* and the control success rate as .50 minus one-half of *r*).

One effect of the standard use of a display procedure such as the Binomial Effect Size Display to index the practical value of our research results would be to give us more useful and more realistic assessments of how well we are really doing as

researchers in the social and behavioral sciences. Employment of the Binomial Effect Size Display has, in fact, shown that we are doing considerably better in our "softer, wilder" sciences than we may have thought we were doing.

So far, our conversation has been intended to comfort the afflicted. In what follows the intent is a bit more to afflict the comfortable. We begin with the topic of replication.

The Meaning of Successful Replication

There is a long tradition in psychology of our urging one another to replicate each other's research. Indeed, there seems to be something nearly scriptural about it--I quote: "If a scholar's work be deemed unreplicable then shall ye gladly cast that scholar out." (That's from either Referees I or Editors II, I believe.)

Now, while we have been very good at calling for replications we have not been too good at deciding when a replication has been successful. The issue we now address is: When shall a study be deemed successfully replicated?

Successful replication is ordinarily taken to mean that a null hypothesis that has been rejected at time 1 is rejected again, and with the same direction of outcome, on the basis of a new study at time 2. The basic model of this usage can be seen in Table 3. The results of the first study are described dichotomously as $p < .05$ or $p >$

Insert Table 3 about here

.05 (or some other critical level, e.g., .01). Each of these two possible outcomes is further dichotomized as to the results of the second study as $p<.05$ or $p>.05$. Thus, cells A and D of Table 3 are examples of failure to replicate because one study was significant and the other was not. Let us examine more closely a specific example of such a "failure to replicate."

Pseudo-Failures to Replicate

The saga of Smith and Jones. Smith has published the results of an experiment in which a certain treatment procedure was predicted to improve performance. She reported results significant at $p<.05$ in the predicted direction. Jones publishes a rebuttal to Smith claiming a failure to replicate.

Insert Table 4 about here

Table 4 shows the results of these two experiments in greater detail. Smith's results were more significant than Jones's, to be sure, but the studies were in perfect agreement as to their estimated sizes of effect as defined either by Cohen's d [(Mean$_1$ – Mean$_2$) / σ] or by r, the correlation between group membership and performance

score (Cohen, 1977; 1988; Rosenthal, 1984). Not only did the effect sizes of the two studies agree, but even the significance levels of .03 and .30 did not differ very significantly: $(Z_{.03} - Z_{.30}) / \sqrt{2} = (2.17 - 1.03) / \sqrt{2} = Z = .81, p = .42$; for details on the comparison of significance levels and effect sizes see Rosenthal and Rubin (1979; 1982a) or a summary in Rosenthal (1984). Table 4 shows very clearly that Jones was very much in error when he claimed that his study failed to replicate that of Smith. Such errors are made very frequently in most areas of psychology and the other behavioral and social sciences. The final column of Table 4 shows that the combined result of both experiments is associated with a more significant *t* and with a smaller confidence interval (for the difference between the means and for the effect size *r*) than is either of the individual studies.

On the odds against replicating significant results. A related error often found in the behavioral and social sciences is the implicit assumption that if an effect is "real," we should therefore expect it to be found significant again upon replication. Nothing could be further from the truth.

Suppose there is in nature a real effect with a magnitude out there in the world of $d = .50$ (i.e., $[\text{Mean}_1 - \text{Mean}_2] / \sigma = .50\ \sigma$ units), or, equivalently, $r = .24$ (a difference in success rate of 62% versus 38%). Then suppose an investigator studies

this effect with an *N* of 64 subjects or so, giving the researcher a level of statistical power of .50, a very common level of power for behavioral researchers of the last 30 years (Cohen, 1962; Sedlmeier & Gigerenzer, in press). Even though a *d* of .50 or an *r* of .24 is a very important effect (as we saw earlier in this paper), there is only one chance in four that both the original investigator and a replicator will get results significant at the .05 level. If there were two replications of the original study there would be only one chance in eight that all three studies would be significant, even though we know the effect in nature is very real and very important.

If five studies investigated this phenomenon, there is only a 50:50 chance that three or more of them would find significant results. In short, given the levels of statistical power at which we normally operate, we have no right to expect the proportion of significant results that we typically do expect, even if in nature there is a very real and very important effect.

Pseudo-Successful Replications

Returning now to Table 3, we focus attention on cell B, the cell of "successful replication." Suppose that two investigators both rejected the null hypothesis at $p < .05$ with both results in the same direction. Suppose further, however, that in one study the effect size *r* was .90 while in the other study the effect size *r* was only .10,

significantly smaller than the *r* of .90 (Rosenthal & Rubin, 1982a). In this case our interpretation is more complex. We have indeed had a successful replication of the rejection of the null hypothesis but we have not come even close to a successful replication of the effect size.

"Successful Replication" of Type 2 Error

Cell C of Table 3 represents the situation in which both studies failed to reject the null hypothesis. Under those conditions investigators might conclude that there was no relationship between the variables investigated. Such a conclusion could be very much in error, the more so as the power of the two studies was low (Cohen, 1977; 1988; Rosenthal, 1986). If power levels of the two studies (assuming medium effect sizes in the population) were very high, say .90 or .95, then two failures to obtain a significant relationship would provide evidence that the effect investigated was not likely to be a very large effect. If power calculations had been made assuming a very small effect size, two failures to reject the null while not providing strong evidence for the null would at least suggest that the size of the effect in the population was probably quite modest.

If sample sizes of the two studies failing to reject the null were small so that power to detect all but the largest effects was low, very little could be concluded from

two failures to reject except that the effect sizes were unlikely to be enormous. For example, two investigators with N's of 20 and 40, respectively, find results not significant at $p<.05$. The effect sizes *phi* (i.e., *r* for dichotomous variables) were .29 and .20, respectively, and both p's were approximately .20. The combined p of these two results, however, is .035 [$(Z_1 + Z_2) / \sqrt{2} = Z$], and the mean effect size in the mid- .20's is not trivial, as we saw earlier in this paper.

Contrasting Views of Replication

The traditional, not very useful view of replication modeled in Table 3 has two primary characteristics:

(1) It focuses on significance level as the relevant summary statistic of a study, and

(2) It makes its evaluation of whether replication has been successful in a dichotomous fashion. For example, replications are successful if both or neither $p<.05$ (or .01, etc.), and they are unsuccessful if one $p<.05$ (or .01, etc.) and the other $p>.05$ (or .01, etc.). Psychologists' reliance on a dichotomous decision procedure accompanied by an untenable discontinuity of credibility in results varying in p levels has been well documented (Nelson, Rosenthal, & Rosnow, 1986; Rosenthal & Gaito, 1963, 1964).

The newer, more useful views of replication success have two primary characteristics:

1. A focus on effect size as the more important summary statistic of a study with only a relatively minor interest in the statistical significance level, and

2. An evaluation of whether replication has been successful made in a continuous fashion. For example, two studies are not said to be successful or unsuccessful replicates of each other, but rather the degree of failure to replicate is specified.

Insert Table 5 about here

Table 5 shows three sets of replications. Replication set A shows two results both rejecting the null but with a difference in effect sizes of .30 in units of *r* or .35 in units of Fisher's *Z* transformation of *r* (Cohen, 1977, 1988; Rosenthal & Rosnow, 1984; Snedecor & Cochran, 1980). That difference, in units of *r* or Fisher's *Z* is the degree of failure to replicate. The fact that both studies were able to reject the null and at exactly the same *p* level is simply a function of sample size. Replication set B shows two studies with different *p* values, one significant at $<.05$, the other not significant. However, the two effect size estimates are in excellent agreement. We would say, accordingly, that replication set B shows more successful replication than

does replication set A. Replication set C shows two studies differing markedly in both level of significance and magnitude (and direction) of effect size. Replication set C, then, is a not very subtle example of a clear failure to replicate.

It should be noted that the values of Table 5 were chosen so that the combined probability of the two studies of sets A, B, and C would all be identical to one another; $(Z_1 + Z_2) / \sqrt{2} = Z$ of 2.77, $p = .0028$, one-tailed.

Some Metrics of the Success of Replication

Once we adopt a view of the success of replication as a function of similarity of effect sizes obtained, we can become more precise in our assessments of the success of replication.

Insert Figure 1 about here

The replication diagonal. Figure 1 shows the "replication plane" generated by crossing the results of the first study conducted (expressed in units of the effect size *r*) by the results of the second study conducted. All perfect replications, those in which the effect sizes are identical in the two studies, fall on a diagonal rising from the lower left corner (–1.00, –1.00) to the upper right corner (+ 1.00, + 1.00). The results of replication set B from Table 5 are shown to fall exactly on the diagonal of perfect

replication (+.26, +.26). The results of replication set A are shown to fallsomewhat above the line representing perfect replication. Figure 1 shows that although set B reflects more successful replication than set A, the latter is also located fairly close to the line and is, therefore, a fairly successful replication set as well. The results of replication set C, however, are shown to fall rather far from the diagonal of perfect replication.

Cohen's q. An alternative to the indexing of the success of replication by the difference between obtained effect size *r*'s is to transform the *r*'s to Fisher's *Z*'s before taking the difference. Fisher's *Z* metric is distributed nearly normally and can thus be used in setting confidence intervals and testing hypotheses about *r*'s, whereas *r*'s distribution is skewed and the more so as the population value of *r* moves further from zero. Cohen's *q* is especially useful for testing the significance of difference between two obtained effect size *r*'s. This is accomplished by means of the fact that

$$q \Big/ \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}$$

is distributed as *Z*, the standard normal deviate (Rosenthal, 1984; Rosenthal & Rubin, 1982a, Snedecor & Cochran, 1980). When there are more than two effect size *r*'s to be evaluated for their variability (i.e., heterogeneity) we can simply compute the standard deviation (*S*) among the *r*'s or their Fisher *Z* equivalents. If a test of

significance of heterogeneity of these Fisher *Z*'s is desired, the three references above all provide the appropriate formula for computing the X^2 test of heterogeneity as does Hedges (1982).

Meta-analytic metrics. As the number of replications for a given research question grows, a full assessment of the success of the replicational effort requires the application of meta-analytic procedures. An informative but slightly unwieldy summary of the meta-analysis might be the stem-and-leaf display of the effect sizes found-in the meta-analysis (Tukey, 1977). A more compact summary of the effect sizes might be Tukey's (1977) box plot, which gives the highest and lowest obtained effect sizes along with those found at the 25th, 50th, and 75th percentiles. For single index values of the consistency of the effect sizes, one could employ (a) the range of effect sizes found between the 75th (Q_3) and 25th (Q_1) percentile, (b) some standard fraction of that range (e.g., half or three-quarters), (c) *S*, the standard deviation of the effect sizes, or (d) SE, the standard error of the effect sizes.

As a slightly more complex index of the stability, replicability, or clarity of the average effect size found in the set of replicates, one could employ the mean effect size divided either by its standard error ($S/\sqrt{k}$ where k is the total number of

replicates), or simply by S. The latter index of mean effect size divided by its standard deviation (S) is the reciprocal of the coefficient of variation or a kind of coefficient of robustness.

The coefficient of robustness of replication. Although the standard error of the mean effect size along with confidence intervals placed around the mean effect size are of great value (Rosenthal & Rubin, 1978), it will sometimes be useful to employ a robustness coefficient that does not increase simply as a function of the increasing number of replications. Thus, if we want to compare two research areas for their robustness, adjusting for the difference in number of replications in each research area, we may prefer the robustness coefficient defined as the reciprocal of the coefficient of variation.

The utility of this coefficient is based on two ideas--first, that replication success, clarity, or robustness depends on the homogeneity of the obtained effect size; and second, that it depends also on the unambiguity or clarity of the directionality of the result. Thus, a set of replications grows in robustness as the variance of the effect sizes decreases and as the distance of the mean effect size from zero increases. Incidentally, the mean may be weighted, unweighted, or trimmed (Tukey, 1977).

Indeed, it need not be the mean at all but any measure of location or central tendency (e.g., the median).

Insert Tables 6 & 7 about here

Table 6 has been prepared to give some feel for the practical meaning of several degrees of variability (S) for seven sets of five replicates each, assuming a mean effect size of zero. For our effect size indicator we have employed the Fisher Z transformation of the correlation coefficient r. When the range of the five Zr's is only from -.02 to +.02, S=.016; when the range is from -1.00 to +1.00, S=.791. Table 7 shows the replication robustness coefficients for each of the seven degrees of variability (S) for each of four levels of mean effect size (Zr): .10, .30, .50, and .70.

There are no intrinsic meanings to any particular robustness coefficients. Instead, they are intended to be used to compare different research domains for their replicational robustness in a merely heuristic way.

What Should Be Reported?

If we are to take seriously our newer view of the meaning of the success of replications, what should be reported by authors of papers seen to be replications of earlier studies? Clearly, reporting the results of tests of significance will not be

sufficient. The effect size of the replication and of the original study must be reported. It is not crucial which particular effect size is employed, but the same effect size should be reported for the replication and the original study. Complete discussions of various effect sizes and when they are useful are available from Cohen (1977, 1988) and elsewhere (e.g., Rosenthal, 1984). If the original study and its replication are reported in different effect size units these can usually be translated to one another (Cohen, 1977, 1988; Rosenthal, 1984; Rosenthal & Rosnow, 1984; Rosenthal & Rubin, in press).

Especially if the results of either the original study or its replication were not significant, the statistical power at which the test of significance was made (assuming, for example, a population effect size equivalent to the effect size actually obtained) should be reported (Cohen, 1988). In addition to reporting the statistical power for each study separately, it would be valuable to report the overall probability that both studies would have yielded significant results given, for example, the effect size estimated from the results of the original and the replication study combined.

As an illustration of this procedure, consider the data of Table 4. Employing Cohen's power tables tells us that given an effect size of $d = .50$, Smith's power to

reject at $p \leq .05$, two-tailed was .60 while Jones's power was .18. Table 8 shows that given these two levels of power there were only 11 chances in a hundred that both studies would reject the null hypothesis given the effect size $d = .50$. Indeed, the odds were three times greater ($p = .33$) that neither study would reject the null hypothesis than that both would reject!

Insert Table 8 about here

Such results are not at all unusual. It has often been documented that behavioral researchers are far fonder of making type II errors than of making type I errors (Cohen, 1962, 1988; Rosenthal & Rosnow, in press; Rosenthal & Rubin, 1985; Sedlmeier & Gigerenzer, in press). It has been suggested that it is part of our Judeo-Christian-Shinto tradition that we be deeply troubled that somewhere out there someone might be having a good time, could be getting a free ride, a significant result they don't deserve, an .05 asterisk that was actually intended for someone else.

A marvelous suggestion has been made by Donald Rubin that would go a long way toward helping us get over our problem with the relative risks of type II versus type I errors. Don has suggested that whenever we conclude that there is "no effect"

we report the effect size along with that confidence interval around the effect size that ranges from the effect size of zero to the equally likely effect size greater than the one we obtained.

To return to Table 4, the "failure to replicate" by Jones provides a good example. Jones did not reject the null but obtained an effect size of $d = .50$. If Jones had been required to report that his d of.50 was just as close to a d of 1.00 as it was to a d of zero, Jones would have been less likely to draw his wrong conclusion.

Meta-Analytic Procedures: An Evaluation

Of course it was bound to happen. No discussion of replication and of the evaluation of the success of a particular replication could long avoid a more formal consideration of meta-analytic procedures.

In the years 1980, 1981, and 1982 alone, well over 300 papers were published on the topic of meta-analysis (Lamb and Whitla, 1983). Does this represent a giant stride forward in the development of the behavioral and social sciences or does it signal a lemming-like flight to disaster? Judging from reactions to past meta-analytic enterprises, there are at least some who take the more pessimistic view. Some three dozen scholars were invited to respond to a meta-analysis of studies of interpersonal expectancy effects conducted by Don Rubin and myself (Rosenthal &

Rubin, 1978). Although much of the commentary dealt with the substantive topic of interpersonal expectancy effects, a good deal of it dealt with methodological aspects of meta-analytic procedures and products. Some of the criticisms offered were accurately anticipated by Glass (1978) who had earlier received commentary on his meta-analytic work (Glass, 1976) and that of his colleagues (Smith & Glass, 1977; Glass, McGaw, & Smith, 1981). In the present discussion, the criticisms of our commentators are grouped into several conceptual categories, described, and discussed.

Sampling Bias and the File Drawer Problem

This criticism holds that there is a retrievability bias such that studies retrieved do not reflect the population of studies conducted. One version of this criticism is that the probability of publication is increased by the statistical significance of the results so that published studies may not be representative of the studies conducted. This is a well-taken criticism, though it applies equally to more traditional narrative reviews of the literature. Procedures that can be employed to address this problem have been described elsewhere (Rosenthal, 1979a; 1984, Chapter 5; Rosenthal & Rubin, 1988).

Loss of Information

Overemphasis on Single Values

The first of two criticisms relevant to information loss notes the danger of trying to summarize a research domain by a single value such as a mean effect size. This criticism holds that defining a relationship in nature by a single value leads to overlooking moderator variables. When meta-analysis is seen as including not only combining effect sizes (and significance levels) but also comparing effect sizes in both diffuse and, especially, in focused fashion, the force of this criticism is removed (Rosenthal, 1984, Chapter 4).

Overlooking negative instances. A special case of the criticism under discussion is that, by emphasizing average values, negative cases are overlooked. There are several ways in which negative cases can be defined; e.g., $p > .05$, $r = 0$, r negative, r significantly negative, and so on. However we may define negative cases, when we divide the sample of studies into negative and positive cases we have merely dichotomized an underlying continuum of effect sizes or significance levels, and accounting for negative cases is simply a special case of finding moderator variables.

Glossing over Details

Although it is accurate to say that meta-analyses gloss over details, it is equally as accurate to say that traditional narrative reviews do so, and that data analysts do so in every study in which any statistics are computed. The act of summarizing requires us to gloss over details. If we describe a nearly normal distribution of scores by the mean and σ we have nearly described the distribution perfectly. If the distribution is quadrimodal, the mean and σ will not do a good job of summarizing the data. It is the data analyst's job in the individual study, and the meta-analyst's job in meta-analysis, to "gloss well.' Providing the reader with all the raw data of all the studies summarized avoids this criticism but serves no useful review function. Providing the reader with a stem-and-leaf display of the effect sizes obtained, along with the results of the diffuse and focused comparisons of effect sizes, does some glossing, but it does a lot of informing besides.

There is, of course, nothing to prevent the meta-analyst from reading each study as carefully and assessing it as creatively as might be done by a more traditional reviewer of a literature. Indeed, we have something of an operational check on reading articles carefully in the case of meta-analysis. If we do not read the results carefully, we cannot obtain effect sizes and significance levels. In traditional

reviews, results may have been read carefully or not read at all, with the abstract or the discussion section providing "the results" to the more traditional reviewer.

Problems of Heterogeneity

Heterogeneity of Method

The first of two criticisms relevant to problems of heterogeneity notes that meta-analyses average over studies in which the independent variables, the dependent variables, and the sampling units are not uniform. How can we speak of interpersonal expectancy effects, meta-analytically, when some of the independent variables are operationalized by (a) telling experimenters that tasks are easy versus hard; or by (b) telling experimenters that subjects are good versus poor task performers? How can we speak, meta-analytically, of these expectancy effects when sometimes the dependent variables are reaction times, sometimes IQ test scores, and sometimes responses to inkblots? How can we speak of these effects when sometimes the sampling units are rats, sometimes college sophomores, sometimes patients, sometimes pupils? Are these not all vastly different phenomena? How can they be pooled together in a single meta-analysis?

Glass (1978) has eloquently addressed this issue--the apples and oranges issue. They are good things to mix, he wrote, when we are trying to generalize to fruit.

Indeed, if we are willing to generalize over subjects within studies, why should we not be willing to generalize over studies? If subjects behave very differently within studies we block on subject characteristics to help us understand why. If studies yield very different results from each other, we block on study characteristics to help us understand why. It is very useful to be able to make general statements about fruit. If, in addition, it is also useful to make general statements about apples, about oranges, and about the differences between them, there is nothing in meta-analytic procedures to prevent us from doing so.

Heterogeneity of Quality

One of the most frequent criticisms of meta-analyses is that "bad" studies are thrown in with good. This criticism must be broken down into two questions: (a) What is a "bad" study?, and (b) What shall we do about "bad" studies?

Defining "bad" studies. Too often, deciding what is a "bad" study is a procedure richly susceptible to bias or to claims of bias (Fiske, 1978). "Bad" studies are too often those whose results we don't like, or, as Glass, McGaw, and Smith (1981) have put it, the studies of our "enemies." Therefore, when reviewers of research tell us they have omitted the "bad" studies, we should satisfy ourselves that this has been done by

criteria we find acceptable. A discussion of these criteria (and the computation of their reliability) can be found elsewhere (Rosenthal, 1984, Chapter 3).

Dealing with "bad" studies. The distribution of studies on a dimension of quality is, of course, not really dichotomous (good versus bad), but continuous with all possible degrees of quality. The fundamental method of coping with "bad" studies or, more accurately, variations in the quality of research, is by differential weighting of studies. Dropping studies is merely the special case of zero weighting.

The most important question to ask relevant to study quality is that asked by Glass (1976): Is there a relationship between quality of research and effect size obtained? If there is not, the inclusion of poorer quality studies will have no effect on the estimate of the average effect size though it will help to decrease the size of the confidence interval around that mean. If there *is* a relationship between the quality of research and effect size obtained, we can employ whatever weighting system we find reasonable (and that we can persuade our colleagues and critics also to find reasonable).

Problems of Independence

Responses Within Studies

The first of two criticisms relevant to problems of independence notes that several effect size estimates and several tests of significance may be generated by the same subjects within each study. This can be a very well-taken criticism under some conditions and the problem has been dealt with elsewhere in some detail (Rosenthal, 1984, Chapter 2; Rosenthal & Rubin, 1986).

Studies Within Sets of Studies

Even when all studies yield only a single effect size estimate and level of significance, and even when all studies employ sampling units that do not also appear in other studies, there is a sense in which results may be nonindependent. That is, studies conducted in the same laboratory, or by the same research group, may be more similar to each other (in the sense of an intraclass correlation) than they are to studies conducted in other laboratories or by other research groups (Jung, 1978; Rosenthal, 1966, 1969, 1979b). The conceptual and statistical implications of this problem are not yet well worked out.

Exaggeration of Significance Levels

Truncating Significance Levels

It has been suggested that all p levels less than .01 (Z values greater than 2.33) be reported as .01 ($Z = 2.33$) because p's less than .01 are likely to be in error (Elashoff, 1978). This truncating of Z's cannot be recommended and will, in the long run, lead to serious errors of inference (Rosenthal & Rubin, 1978). If there is reason to suspect that a given p level $< .01$ is in error it should, of course, be corrected before employing it in the meta-analysis. It should not, however, be changed to $p = .01$ simply because it is less than .01.

Too Many Studies

It has been noted as a criticism of meta-analyses, that, as the number of studies increases, there is a greater and greater probability of rejecting the null hypothesis (Mayo, 1978). When the null hypothesis is false and, therefore, ought to be rejected, it is indeed true that adding observations (either sampling units within studies or new studies) increases statistical power. However, it is hard to accept as a legitimate criticism of a procedure, a characteristic that increases its accuracy and decreases its error rate-- in this case, type II errors.

A related feature of meta-analysis appears to be that it may, in general, lead to a decrease in type II errors even when the number of studies is modest. Empirical support for this is provided in a study conducted by Cooper and Rosenthal (1980). Procedures requiring the research reviewer to be more systematic and to use more of the information in the data seem to be associated with increases in power, i.e., decreases in type II errors.

Some Benefits of Meta-Analysis

From what has been said of the various criticisms of meta-analysis it will surprise no one to learn that I strongly support the increasing use of meta-analytic procedures. My reasons for that support go beyond the fact that the various criticisms of meta-analysis can be readily addressed. In the time that remains I want to note a number of special benefits of meta-analysis. Some of these benefits are well known, but some are not--indeed, some are almost secret benefits.

Most Obvious Benefits

Completeness. Meta-analytic consideration of a research domain is more complete and exhaustive though this does *not* mean that all studies found are weighted equally. Indeed, every study should be weighted from zero to any desired

number. These weights, of course, must be defensible. (It will not do to weight all my results + 1.00 and all my enemies' results 0.00).

Explicitness. The quantitative nature of the process of obtaining effect sizes, standard normal deviates, and weights, forces explicitness on the analyst. Vague terms like "no relationship," "some relationship," a "strong relationship," "very significant," are replaced by numerical values.

Power. Empirical work has shown that meta-analytic procedures increase power and decrease type 2 errors (Cooper & Rosenthal, 1980).

Less Obvious Benefits

Moderator variables. These are more easily spotted and evaluated in a context of a quantitative research summary. This aids theory development and increases empirical richness.

Cumulation problems. Meta-analytic procedures address, in part, the chronic complaint that social sciences cumulate so poorly compared to the physical sciences. It should be noted that recent historical and sociological investigations have suggested that the physical sciences may not be all that much better off than we are when it comes to successful replication (Collins, 1985; Hedges, 1987; Pool 1988). For example, Collins (1985) has described the failures to replicate the construction of

TEA-lasers despite the availability of detailed instructions for replication. Apparently TEA-lasers could be replicated dependably only when the replication instructions were accompanied by a scientist who had actually built a laser.

Least Obvious Benefits

Decrease in overemphasis on single studies. One not so obvious benefit that will accrue to us is the gradual decrease in the overemphasis on the results of a single study. There are good sociological grounds for our monomaniacal preoccupation with the results of a single study. Those grounds have to do with the reward system of science where recognition, promotion, reputation, and the like depend on the results of the single study, also known as the smallest unit of academic currency. The study is "good," "valuable," and above all, "publishable" when $p \leq .05$. Our disciplines would be further ahead if we adopted a more cumulative view of science in which the impact of a study were evaluated less on the basis of p levels, and more on the basis of its own effect size and on the revised effect size and combined probability that resulted from the addition of the new study to any earlier studies investigating the same or a similar relationship. This, of course, amounts to a call for a more meta-analytic view of "doing science."

B. F Skinner has been eloquent in his comments on the overvaluation of the single study: "In my own thinking, I try to avoid the kind of fraudulent significance which comes with grandiose terms or profound 'principles.' But some psychologists seem to need to feel that every experiment they do demands a sweeping reorganization of psychology as a whole. It's not worth publishing unless it has some such significance. But research has its own values, and you don't need to cook up spurious reasons why it's important." (Skinner, 1983, p. 39).

"The new intimacy." This new intimacy is between the reviewer and the data. We cannot do a meta-analysis by reading abstracts and discussion sections. We are forced to look at the numbers and, very often, compute the correct ones ourselves. Meta-analysis requires us to cumulate *data*, not *conclusions*. "Reading" a paper is quite a different matter when we need to compute an effect size and a fairly precise significance level--often from a results section that never heard of effect sizes, precise significance levels (or the APA publication manual)!

The demise of the dichotomous significance testing decision. Far more than is good for us, social and behavioral scientists operate under a dichotomous null hypothesis decision procedure in which the evidence is interpreted as anti-null if $p \leq .05$ and pro-null if $p > .05$. If our dissertation p is $< .05$ it means joy, a Ph.D., and a

tenure-track position at a major university. If our p is $> .05$ it means ruin, despair, and our advisor's suddenly thinking of a new control condition that should be run. That attitude really must go. God loves the .06 nearly as much as the .05. Indeed, I have it on good authority that she views the strength of evidence for or against the null as a fairly continuous function of the magnitude of p. As a matter of fact, two .06 results are much stronger evidence against the null than one .05 result; and 10 p's of .10 are stronger evidence against the null than 5 p's of .05.

The overthrow of the omnibus test. It is common to find specific questions addressed by F tests with $df > 1$ in the numerator or by X^2 tests with $df > 1$. For example, suppose the specific question is whether increased incentive level improves the productivity of work groups. We employ four levels of incentive so that our omnibus F test would have 3 df in the numerator or our omnibus X^2 would be on at least 3 df. Common as these tests are, they reflect poorly on our teaching of data analytic procedures. The diffuse hypothesis tested by these omnibus tests usually tells us nothing of importance about our research question. The rule of thumb is unambiguous: Whenever we have tested a fixed effect with $df > 1$ for X^2 or for the numerator of F, we have tested a question in which we are almost surely not interested.

The situation is even worse when there are several dependent variables as well as multiple *df* for the independent variable. The paradigm case here is canonical correlation and special cases are MANOVA, MANCOVA, Multiple discriminant function, multiple path analysis, and complex multiple partial correlation. While all of these procedures have useful exploratory data analytic applications they are commonly used to test null hypotheses which are scientifically almost always of doubtful value. The effect size estimates they yield (e.g., the canonical correlation) are also almost always of doubtful value.

This is not the place to go into detail, but one approach to the problem of analyzing canonical data structures is to reduce the set of dependent variables to some smaller number of composite variables using the principal-components-followed-by-unit-weighting approach. Each composite can then be analyzed serially.

Meta-analytic questions are basically contrast questions. F tests with $df>1$ in the numerator or χ^2's with $df>1$ are useless in meta-analytic work. That leads to an additional scientific benefit:

The increased recognition of contrast analysis. Meta-analytic questions require precise formulation of questions and contrasts are procedures for obtaining answers to such questions, often in an analysis of variance or table analysis context.

Although most textbooks of statistics describe the logic and the machinery of contrast analyses, one still sees contrasts employed all too rarely. That is a real pity given the precision of thought and theory they encourage and (especially relevant to these times of publication pressure) given the boost in power conferred with the resulting increase in .05 asterisks (Rosenthal & Rosnow, 1985).

A probable increase in the accurate understanding of interaction effects. Probably the universally most misinterpreted empirical results in psychology are the results of interaction effects. A recent survey of 191 research articles involving interactions found only two articles that showed the authors interpreting interactions in an unequivocally correct manner (i.e., by examining the residuals that define the interaction) (Rosnow & Rosenthal, 1989). The rest of the articles simply compared means of conditions with other means, a procedure that does not investigate interaction effects but rather the sum of main effects and interaction effects.

Most standard textbooks of statistics for psychologists provide accurate mathematical definitions of interaction effects but then interpret not the residuals that define those interactions but the means of cells that are the sums of all main effects and all interactions.

In addition, users of SPSS, SAS, BMDP, and virtually all other data-analytic software are poorly served in the matter of interactions since virtually no programs provide convenient tabular output giving the residuals defining interaction. The only exception to that of which I am aware is a little-known package called Data-Text developed by Arthur Couch and David Armor for which William Cochran and Donald Rubin provided the statistical consultation.

Since many meta-analytic questions are by nature questions of interaction (for example, that opposite sex dyads will conduct standard transactions more slowly than will same sex dyads), we can be hopeful that increased use of meta-analytic procedures will bring with it increased sophistication about the meaning of interaction.

Meta-analytic procedures are applicable beyond meta-analyses. Many of the techniques of contrast analyses among effect sizes, for example, can be used within a single study (Rosenthal & Rosnow, 1985). Computing a single effect size from correlated dependent variables, or comparing treatment effects on two or more dependent variables serve as illustrations (Rosenthal & Rubin, 1986).

The decrease in the splendid detachment of the full professor. Meta-analytic work requires careful reading of research and moderate data analytic skills. We

cannot send an undergraduate research assistant to the library with a stack of 5×8 cards to bring us back "the results." With narrative reviews that seems often to have been done. With meta-analysis the reviewer must get involved with the actual data and that is all to the good.

Conclusion

I hope that this paper has provided some comfort to the afflicted in showing that many of the findings of our discipline are neither as small nor as unimportant from a practical point of view as we may have feared. Perhaps I hope, too, that there may have been some affliction of the comfortable in showing that in our views of replication and of the cumulation of the wisdom of our field there is much yet remaining to be done.

References

Barnes, D. M. (1986). Promising results halt trial of anti-AIDS drug. *Science, 234*, 15-16.

Canadian Multicentre Transplant Study Group. (1983). A randomized clinical trial of cyclosporine in cadaveric renal transplantation. *New England Journal of Medicine, 309*, 809-815.

Centers for Disease Control Vietnam Experience Study. (1988). Health status of Vietnam veterans: 1. Psychosocial characteristics. *Journal of the American Medical Association, 259*, 2701-2707.

Cohen, J. (1962). the statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145-153.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed., Hillsdale, NJ: Lawrence Erlbaum Associates.

Collins, H. M. (1985). *Changing Order: Replication and Induction in Scientific Practice*. Beverly Hills, CA: Sage.

Cooper, H. M., & Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin, 87*, 442-449.

Elashoff, J. D. (1978). Box scores are for baseball. *The Behavioral and Brain Sciences*, *3*, 392.

Fiske, D. W. (1978). The several kinds of generalization. *The Behavioral and Brain Sciences*, *3*, 393-394.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*, 3-8.

Glass, G. V. (1978). In defense of generalization. *The Behavioral and Brain Sciences*, *3*, 394-395.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-Analysis in Social Research*. Beverly Hills, CA: Sage.

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, *92*, 490-499.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? *American Psychologist*, *42*, 443-455.

Jung, J. (1978). Self-negating functions of self-fulfilling prophecies. *The Behavioral and Brain Sciences*, *3*, 397-398.

Lamb, W. K., & Whitla, D. K. (1983). *Meta-Analysis and the Integration of Research Findings: A Trend Analysis and Bibliography Prior to 1983*. Unpublished manuscript, Harvard University, Cambridge.

Mayo, R. J. (1978). Statistical considerations in analyzing the results of a collection of experiments. *The Behavioral and Brain Sciences, 3*. 400-401.

Nelson, N., Rosenthal, R., & Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist, 41*. 1299-1301.

Pool, R. (1988). Similar experiments, dissimilar results. *Science, 242*. 192-193.

Rimland, B. (1979). Death knell for psychotherapy? *American Psychologist, 34*. 192.

Rosenthal, R. (1966). *Experimenter Effects in Behavioral Research*. New York: Appleton-Century-Crofts.

Rosenthal, R. (1969). Interpersonal expectations. In R. Rosenthal and R. L. Rosnow (Eds.), *Artifact in Behavioral Research* (pp. 181-277). New York: Academic Press.

Rosenthal, R. (1979a). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*, 638-641.

Rosenthal, R. (1979b). Replications and their relative utilities. *Replications in Social Psychology, 1*(1), 15-23.

Rosenthal, R. (1984). *Meta-Analytic Procedures for Social Research*. Beverly Hills, CA: Sage.

Rosenthal, R. (1986). Nonsignificant relationships as scientific evidence. *Behavioral and Brain Sciences, 9*, 479-481.

Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology, 55*, 33-38.

Rosenthal, R., & Gaito, J. (1964). Further evidence for the cliff effect in the interpretation of levels of significance. *Psychological Reports, 15*, 570.

Rosenthal, R., & Rosnow, R. L. (1984). *Essentials of Behavioral Research: Methods and Data Analysis*. New York: McGraw-Hill.

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast Analysis: Focused Comparisons in the Analysis of Variance*. New York: Cambridge University Press.

Rosenthal, R., & Rosnow, R. L. (in press). *Essentials of Behavioral Research: Methods and Data Analysis*. 2nd ed., New York: McGraw-Hill.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *The Behavioral and Brain Sciences, 3*, 377-386.

Rosenthal, R., & Rubin, D. B. (1979). Comparing significance levels of independent studies. *Psychological Bulletin, 86*, 1165-1168.

Rosenthal, R., & Rubin, D. B. (1982a). Comparing effect sizes of independent studies. *Psychological Bulletin, 92*, 500-504.

Rosenthal, R., & Rubin, D. B. (1982b). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74*, 166-169.

Rosenthal, R., & Rubin, D. B. (1985). Statistical analysis: Summarizing evidence versus establishing facts. *Psychological Bulletin, 97*, 527-529.

Rosenthal, R., & Rubin, D. B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin, 99*, 400-406.

Rosenthal, R., & Rubin, D. B. (1988). Comment: Assumptions and procedures in the file drawer problem. *Statistical Science, 3*, 120-125.

Rosenthal, R., & Rubin, D. B. (in press). Effect size estimation for one-sample multiple-choice-type data: Design, analysis, and meta-analysis. *Psychological Bulletin*.

Rosnow, R. L., & Rosenthal, R. (1989). Definition and interpretation of interaction effects. *Psychological Bulletin, 105*, 143-146.

Sedlmeier, P., & Gigerenzer, G. (in press). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*.

Skinner, B. F. (1983, August). On the value of research. *APA Monitor*, p. 39.

Smith, M. L., & Glass, G. V (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32*, 752-760.

Snedecor, G. W., & Cochran, W. G. (1980). *Statistical Methods* (7th ed.). Ames: Iowa State University Press.

Steering Committee of the Physicians Health Study Research Group. (1988). Preliminary report: Findings from the aspirin component of the ongoing physicians' health study. *The New England Journal of Medicine, 318*, 262-264.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Author Notes

This paper was presented as an EPA Distinguished Lecture at the Annual Meeting of the Eastern Psychological Association, Boston, April 2, 1989. Preparation of this paper was supported in part by the National Science Foundation while the author was a Fellow at the Center for Advanced Study in the Behavioral Sciences. I am grateful for financial support provided by the John D. and Catherine T. MacArthur Foundation, and for improvements suggested by Lynn Gale, Deanna Knickerbocker, Harold Luft, and Lincoln Moses.

Table 1

Effects of Aspirin on Heart Attacks Among 22,000 Physicians

	Heart Attack	No Heart Attack	Total
I. *Raw Counts*			
Aspirin	104	10,933	11,037
Placebo	189	10,845	11,034
Total	293	21,778	22,071
II. *Percentages*			
Aspirin	0.94	99.06	100
Placebo	1.71	98.29	100
Total	1.33	98.67	100
III. *Binomial Effect Size Display*			
Aspirin	48.3	51.7	100
Placebo	51.7	48.3	100
Total	100	100	200

Table 2

Other Examples of Binomial Effect Size Displays

I. *Vietnam Service and Alcohol Problems* (r = .07)

	Problem	No Problem	Total
Vietnam Veteran	53.5	46.5	100
Non-Vietnam Veteran	46.5	53.5	100
Total	100	100	200

II. *AZT in the Treatment of AIDS* (r = .23)

	Death	Survival	Total
AZT	38.5	61.5	100
Placebo	61.5	38.5	100
Total	100	100	200

III. *Benefits of Psychotherapy* (r = .32)

	Less Benefit	Greater Benefit	Total
Psychotherapy	34	66	100
Control	66	34	100
Total	100	100	200

Table 3

Common Model of Successful Replication: Judgment is Dichotomous and Based on Significance Testing

		First Study	
		$p > .05$*	$p < .05$
Second Study	$p < .05$[a]	A Failure to Replicate	B Successful Replication
	$p > .05$	C Failure to Establish Effect	D Failure to Replicate

*By convention .05 but could be any other given level, e.g. .01.

[a]In the same tail as the results of the first study.

Table 4

Illustrative Results of an Experiment and Its Replication

		Investigator		
		I. *Smith*	II. *Jones*	Combined
Treatment Mean		.38	.36	.376
Control Mean		.26	.24	.256
Difference		.12	.12	.120
t		2.21	1.06	2.45
df		78	18	96
two-tail *p*		.03	.30	.02
effect size *d*[a]		.50	.50	.50
effect size *r*[b]		.24	.24	.24
standard normal *Z*		2.17[c]	1.03[c]	2.40
95% Confidence intervals				
Mean differences	From:	.01	-.12	.02
	To:	.23	.36	.22
Effect size *r*'s	From:	.02	-.23	.04
	To:	.44	.62	.42

[a] Obtained from $2t\sqrt{df}$.

[b] Obtained from $\sqrt{t^2/(t^2+df)}$.

[c] These significance levels differ at $Z = .81$, $p = .42$ from $(Z_1 - Z_2)/\sqrt{2}$.

Table 5

Comparison of Three Sets of Replications

	Replication Sets					
	A		*B*		*C*	
	Study 1	*Study 2*	*Study 1*	*Study 2*	*Study 1*	*Study 2*
N	96	15	98	27	12	32
p (two-tail)	.05	.05	.01	.18	.000001	.33
Z(*p*)	1.96	1.96	2.58	1.34	4.89	-0.97
r	.20	.50	.26	.26	.72	-.18
Z(*r*)	.20	.55	.27	.27	.90	-.18
Cohen's $q\,(Z_{r_1}-Z_{r_2})$	.35		.00		1.08	

Table 6

Seven Degrees of Variability (*S*) of Effect Sizes (*Zr*) Around a Mean Effect Size of 0.00

	Degree of Variability						
Replicate	*Set 1*	*Set 2*	*Set 3*	*Set 4*	*Set 5*	*Set 6*	*Set 7*
	.02	.10	.20	.40	.60	.80	1.00
	.01	.05	.10	.20	.30	.40	.50
	.00	.00	.00	.00	.00	.00	.00
	-.01	-.05	-.10	-.20	-.30	-.40	-.50
	-.02	-.10	-.20	-.40	-.60	-.80	-1.00
S	.016	.079	.158	.316	.474	.632	.791
Range	.04	.20	.40	.80	1.20	1.60	2.00
Equal Steps of	.01	.05	.10	.20	.30	.40	.50

Table 7

Replication Robustness Coefficients for Four Levels of Mean Effect Size (Zr) and Six Degrees of Variability of Effect Size (S)

	Mean Effect Size (Zr)			
S	.10	.30	.50	.70
.016	6.25	18.75	31.25	43.75
.079	1.27	3.80	6.33	8.86
.158	0.63	1.90	3.16	4.43
.316	0.32	0.95	1.58	2.22
.474	0.21	0.63	1.05	1.48
.632	0.16	0.47	0.79	1.11
.791	0.13	0.38	0.63	0.88

Table 8

Probabilities of Various Combinations of Rejecting the Null Hypothesis for the Two Studies of Table 4

	Study I: Smith		
	Probability of Not Rejecting False Null (Type II Error Rate = .40)	Probability of Rejecting False Null (Power = .60)	Σ
Study II: Jones			
Probability of Rejecting False Null (Power = .18)	.07	.11	.18
Probability of Not Rejecting False Null (Type II Error Rate = .82)	.33	.49	.82
Σ	.40	.60	1.00

Figure 1

The Replication Plane